FE Project Summary 09/11/2013

## Predict Price

* Summary

We used two methods to parse the data and try to figure out the relationship between price change and SIO.

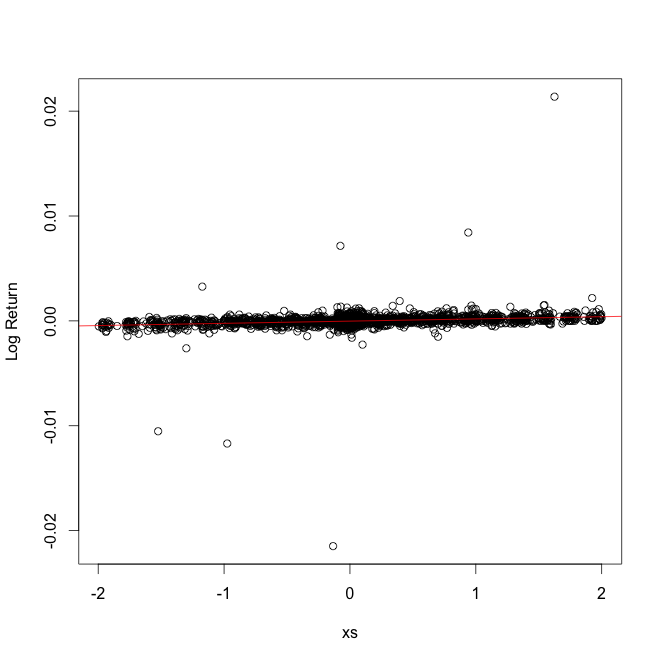
So far, the results are weak. Those weak results may be related to the parameters we choose like smaller time bin size, bucket size, threshold value, the form of the variables, and etc. In the future analysis, we may consider stock with larger volume, or use function to test different combination of a series of the parameters and use transformation of the variables, i.e. power form, exponential form, log form and so on. Also more independent variables can be included, like exponential weighted IO.

In the further analysis, I want to separate the model into two parts. The first is to use categorical analysis to predict price change direction, and the second to predict the change magnitude. In the categorical regression part, assign three directions to the price change, going-up, unchanged, and going-down. There will also be a new parameter involved here (threshold to define the three directions), and needed to use automated function to select the best one. Moreover, as I know, two logistic regression may produce stronger result than one categorical analysis, which we can further look into.

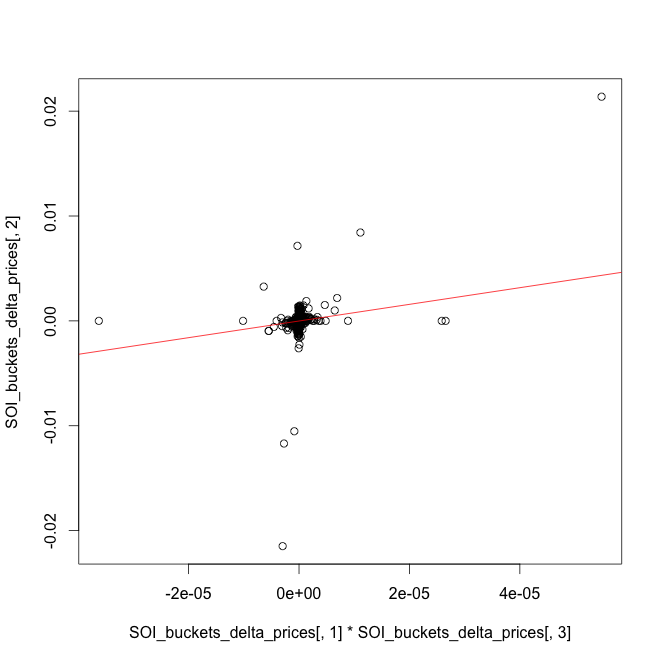
1. Predict Price with SIO

* Weak relationship between the SIO and log price return

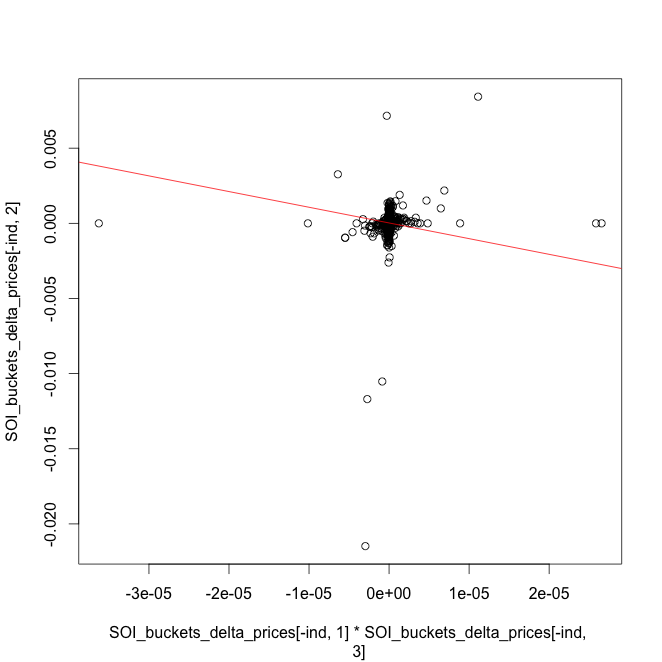
In order to see if SIO has a relationship with price return, I first regress the log return to SIO calculated from trade data. The relationship is not strong with R^2 = 0.04.



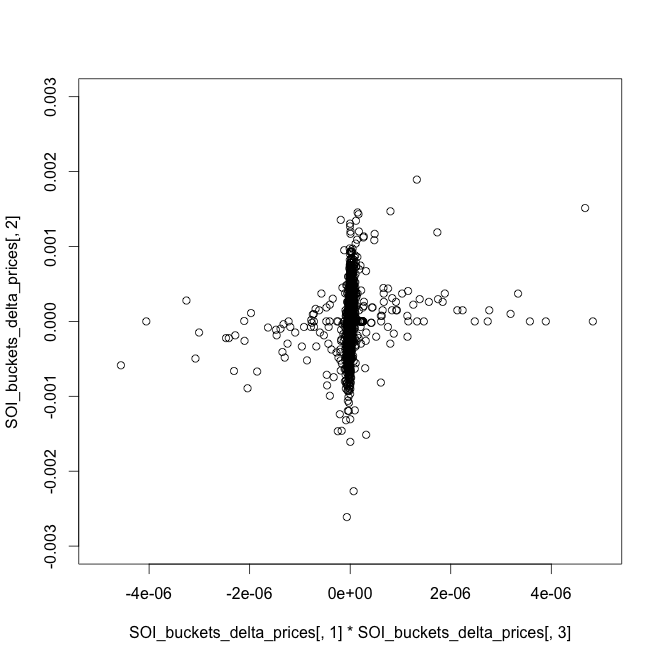
Since Zhenyu also calculate the bucket log price volatility, I think the volatility may have a combined effect on the SIO. So I add the cross product of these two variables as a new independent variable in the regression. The R^2 increased to 22%. Take a closer look at the additional predict power, it is due to an influential point lies in the upper right corner in the plot below.



Take this point out, the R^2 decreased to 0.095, and the plot is as following:

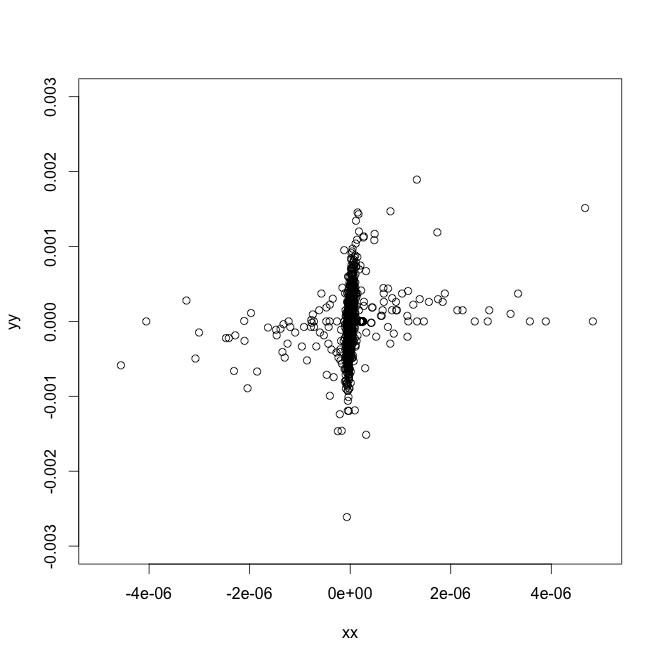


There are still some influential points around. Zoom in the crowed part to see a clearer pattern:

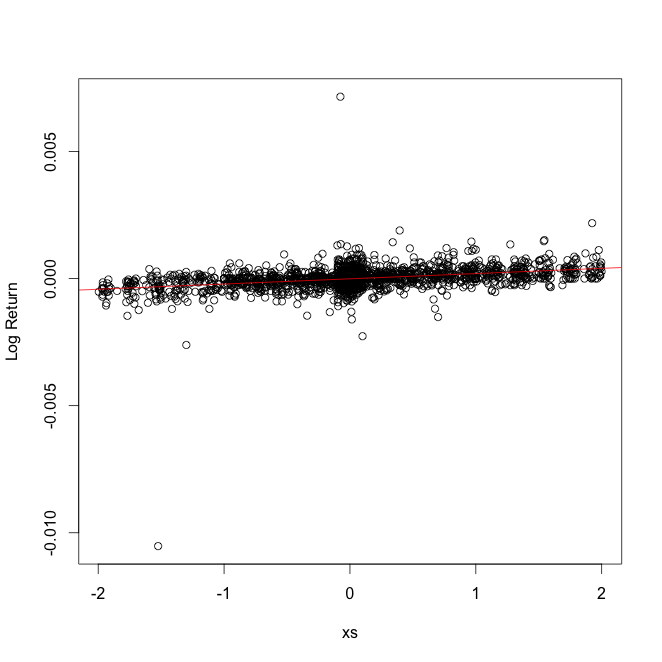


Lots points are concentrated in the vertical line, which means the cross product variable has many values near zero. We think SIO around zero is the most cases and should be not indicative for the price change, we first exclude those points to see if there is some good linear trend left.

We exclude points that either the IO is smaller than 0.3, or volatility equals zero. However, the preliminary result is not positive. The pattern remains the same as in the plot below.



Another interesting finding is that when I exclude the last 200+ samples (just use the first 2000 buckets), of which the log price change is mostly zero, to urn the regression: log price change ~ SIO, the results got improved (R^2 12.5%). This result can be further improved to 14% when I add the cross product term.



* Prediction

Since the relationship is very weak, we expect the predict power will be even weaker.

Based on the last regression I change the independent variable to a bucket ahead. The R^2 is 0.01.

1. Predict price with SIO calculated from quotes and trades data

Use quotes to assign BUY or SELL to each trade;

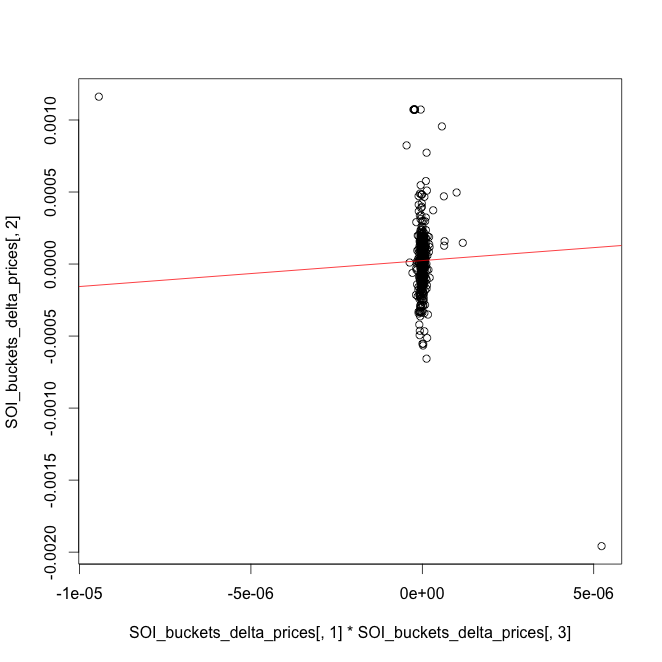
Price change = log(Instantaneous Price) – SMA(log price)

The analysis is conducted the same way as in part 1.

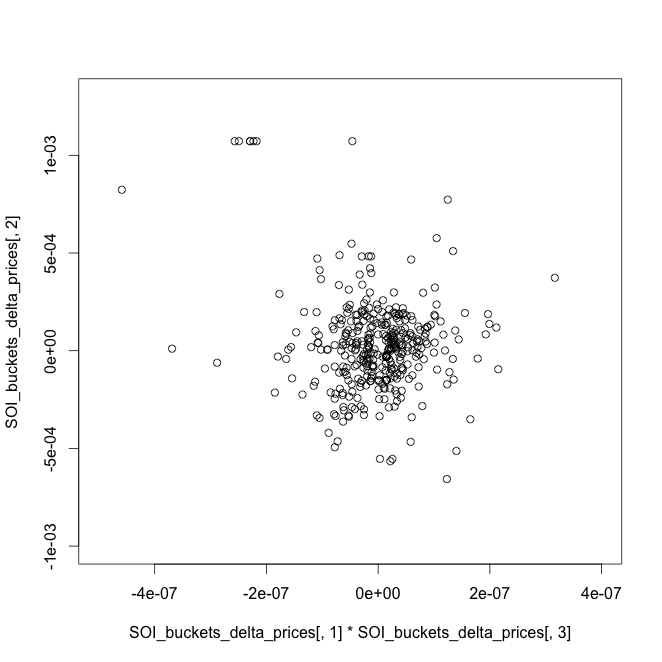
* Weak relationship between the SIO and log price return

R^2 equals 0.0025%

Add the product of volatility and SIO as a new independent variable, the R^2 improved to 0.16. This result has two influential points as in the plot below.

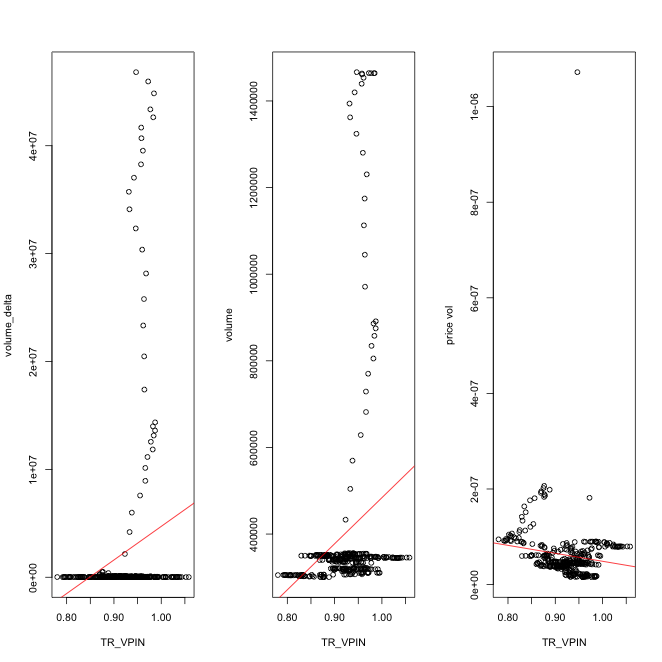


Zoom in, I get:



No obvious linear trend in this plot as well.

## VPIN predict crash



Off-set by 1 period

